所属类别	**2020 年“华数杯”全国大学生数学建模竞赛**	参赛编号

脱贫帮扶绩效评价与建议

摘要

扶贫作为党的富民政策的组成部分，作为推动农村经济体制深入改革的内容，越来越受到全党的重视和全社会的支持。为了更好的激励各帮扶单位更好的扶贫，我国政府启动脱贫帮扶评价机制，从多方面多角度对各扶贫单位进行绩效评价。

对于问题一，为了恢复数据的客观真实性以便将来得到更好的分析结果,我们使用拉伊达方法（非等置信概率）剔除异常值。如果某测量值与平均值只差大于标准偏差的三倍，则予以剔除。另外，我们使用“加权移动平均”平滑滤波对数据进行平滑处理。为了探究各个评价指标之间的对应关系，采用相关分析的方法进行研究。2015 年的评分与 2020 年对应的各项评分之间散点图可以简单看出对应变量之间的线性相关关系，进一步利用 Spearman 秩和相关系数进行检验，结果显示：在显著性水平为 0.01 时，检验的 p 值均小于 0.01，通过显著性检验，即 5 组对应评分之间存在显著线性关系。相关系数 r 的取值满足 $0.5 \le r \le 0.8$，说明 5 组对应变量之间存在中度相关关系。

对于问题二，①以帮扶单位类型(0-5)作为划分依据，对脱贫帮扶绩效排序。以 6 个帮扶类型总分的相对增长量(Y)作为评价因素，选用各组均值作为检验参数，进行单因素方差分析及非参数方法 Jonckheere-Terpstra 单侧检验。两种检验结果相同，结果显示：帮扶单位类型绩效由高到低依次为 5、0、2、1、3、4。②以帮扶单位编号作为划分依据，对脱贫帮扶绩效排序。以 160 个帮扶单位的平均相对增长量的五项指标作为评价因素，建立 Topsis 模型。以熵值法确定指标权重，计算正负理想解与综合得分。计算得出：单位编号为 142、60、74 、137、144 、112 、158 、76、147 和 114 为脱贫帮扶绩效排名前十名。

对于问题三，考虑到每一帮扶单位的工作特点，需要继续探究各个因素的排序情况。首先构建评价指标体系，在原始数据基础上纳入所有衡量相对增长量的指标建立因子分析模型。通过因子分析，指定提取因子个数并进行旋转命名，依据因子得分分别进行排序，并将每一项因子的排名前五名进行归纳与分析。最后，提出模型的改进与思考建议。结果表明：模型适用型良好且结果较为合理。

对于问题四，以帮扶单位（0-159）作为研究对象，在问题二评价模型排序的基础上，以评为脱贫先进村庄的比例对数据划分为先进村庄（1）和其余（0）两类作为因变量，五个指标的相对增量为自变量，建立 Logistic 回归模型，并利用混淆矩阵和 ROC 曲线等评价指标进行评价。结果显示：居民收入相对增量指标对脱贫先进村庄评级影响力度最大，模型准确度达到 0.946，适用性良好。并运用结论和帮扶单位的类型与编号进行类比求均值，对其余十组数据的 2020 年缺失值进行填补，带入方程计算 y 并排序，将村庄编号为 22096 评为脱贫先进一级村庄，34208，52436 和 47883 评为二级村庄。

对于问题五，我们针对贫困地区贫困面广、基础设施薄弱、受教育水平低、产业化落后以及扶贫资金不足等问题，针对性提出了加强对口帮扶责任制、加强基础设施建设扶贫、实施精准扶贫、加大造血式扶贫、创新政策式扶贫等五个解决方案。

关键词：数据处理；相关分析；J-T 检验；Topsis；因子分析；二元 Logistic

1 问题重述

1.1 问题背景

贫富分化是一道世界性的发展难题。消除贫困，缩小贫富差距，不仅是中国政府多年来致力解决的难题，同时也是世界许多国家与地区所面临的共同难题。习近平总书记曾多次强调，“扶贫要实事求是、因地制宜、分类指导、精准扶贫”。这些年来，我国的扶贫工作取得了举世瞩目的伟大成就，探索出了扶贫开发的中国道路。但贫困人口仍然占到很大一个比例，扶贫工作还存在诸多不如人意的地方。

为了解决精准扶贫中所存在的问题，早日实现共同富裕，我国政府启动脱贫帮扶绩效评价机制，从多方面多角度对各扶贫单位进行绩效评价，以此激励各帮扶单位更好的扶贫，提高脱贫效率、扶真贫、真扶贫[1]。

1.2 问题提出

问题一，针对五年前的评分与 2020 年对应的各项评分是否有直接的关联提出观点，并分析各个指标的对应关系。

问题二，考虑到不同帮扶单位在各方面存在差异，仅用 2020 年各村庄评分高低无法有效的体现一个帮扶单位在脱贫攻坚提升方面所做出的努力。运用附件数据，阐明什么类型的帮扶单位，哪些帮扶单位在脱贫帮扶上面有较高的绩效？给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号。

问题三，每个帮扶单位在扶贫上有不同的工作特色。请给出，哪些帮扶单位分别在居民收入、产业发展、居住环境、文化教育、基础设施等评价指标上帮扶业绩明显？请列出各单项评价指标前五名的帮扶单位编号。

问题四，全国计划给与 10000 个村庄“脱贫先进村庄”称号。哪些因素对获得这个荣誉称号有着非常重要的影响？并对数据表中最后的 10 个村庄数据填补与判断是否能评上“脱贫先进村庄”称号。若一、二级称号比例为 1:3，则 10 个村庄中谁能评上“脱贫先进一级村庄”称号？

问题五，根据研究成果，向国家扶贫办写一封 500 字左右的信，阐述观点和建议。

2 问题分析

2.1 总体分析

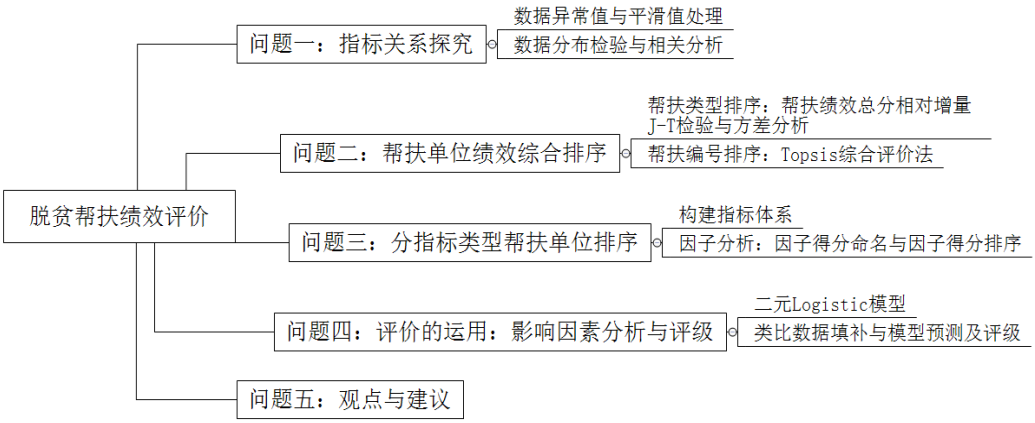


图 1 文章框架图

2.2 具体分析

针对问题一，由于测量数据在其采集与传输过程中，由于环境干扰或人为因素有可能造成个别数据不切合实际或丢失，即出现异常值。为了恢复数据的客观真实性以便将来得到更好的分析结果,有必要先对原始数据剔除异常值。另外，无论是人工观测的数据还是由数据采集系统获取的数据，都不可避免叠加上“噪声”干扰，反映在曲线图形上就是一些“毛刺和尖峰”。为了提高数据的质量，必须对数据进行平滑处理去噪声干扰。为了验证一般的理解思路——5 年前后对应的各项评分有直接的关联，并分析各个评价指标之间的对应关系，一般会采用相关分析的方法进行研究。相关分析的种类有很多种，可以先对 5 年前后对应的各项评分之间的散点图进行观察，初步满足线性相关关系，进一步利用 Spearman 秩和相关系数进行检验，若通过显著性检验，即证实 5 组对应评分之间存在显著线性关系。通过相关系数 r 的取值大小，判断 5 组对应变量之间相关关系强弱。

针对问题二，①首先需要阐明什么类型的帮扶单位在脱贫帮扶上有较高绩效，所以考虑以帮扶单位类型(0-5)作为划分依据，一般可以采用方差分析及非参数方法 Jonckheere-Terpstra 单侧检验来进行评价。选取 6 个帮扶类型总分的相对增长量(Y)作为评价因素，以各组均值作为检验参数，进行两种方法的检验，两种检验结果相同。结果显示：帮扶单位类型绩效由高到低依次为 5、0、2、1、3、4。②帮扶单位编号排序问题仍然采用五个指标的相对增长量作为主要的参考指标。这是由于仅仅从 2020 年的数据或者是单纯 2015 年的数据无法体现出帮扶绩效。因此，以这 160 个帮扶单位的每一单位的平均相对增长量的五项指标作为评价因素，利用评价模型中最为常见的 Topsis 模型进行建模。首先根据熵值法（数据的变异程度）确定指标的权重，在此基础上计算正负理想解并计算综合得分。以综合得分对 160 个单位的数据进行排序，计算结果得出：扶贫单位编号为 142、60、74 、137、144 、112 、158 、76、147 和 114 为排名前十。

针对问题三，在对帮扶单位按不同类型和按不同编号的总体绩效评价进行分析之后，考虑到每一帮扶单位有所使侧重点与特点。因此，需要继续探究各个因素的排序情况。首先构建评价指标体系，选择 2015 和 2010 年的五个指标的原始数据纳入评价体系中，同时加入总分的相对增长量，共计 16 个指标建立因子分析模型。试图通过因子分析，指定因子提取个数以对每一个因子尽可能的再次对其五个方面进行命名。在因子旋转后对其因子得分进行排序，得出结果。其结果表明：因子分析适用型良好，帮扶单位 1 号在改善居住环境上帮助的效果最好，1149 号在改善文化教育上帮助效果最好，115 号在提高居民收入帮助效果最好，106 号在改善基础建设上帮助效果最好，帮扶单位 72 号在提高产业帮助上的效果最好。最后，提出模型的改进与思考建议。

针对问题四，以问题二所研究的编号 0 至 159 帮扶单位作为研究对象，根据其评为脱贫先进村庄的比例关系，对这 160 个帮扶单位进行处理后划分为先进村庄（1）和其余（0）两大类作为二元因变量，其余五个指标的增长相对量为自变量，建立 Logistic 回归模型。根据各个系数大小对指标对评级的影响力进行排序与分析，并利用混淆矩阵和 ROC 曲线等评价指标进行模型的评价。结果显示：居民收入相对增长量指标对脱贫先进村庄评级影响力度最大，具有重要影响，模型准确度（auc）达到 0.946，模型适用性良好。在此基础上运用结论和村庄的帮扶单位的类型与编号进行类比求均值，对其余十组数据的 2020 年指标值进行填补，并计算增长量带入模型方程计算 y，根据概率大小排序，可以将村庄编号为 22096 评价为脱贫先进一级村庄，村庄编号为 34208，52436 和 47883 评为二级村庄。

针对问题五，目前贫困地区的生存发展状况不容乐观，主要表现为五个方面。一是贫困面广、贫困人口多、贫困程度深；二是基础设施薄弱，成为贫困地区经济发展的瓶

颈；三是受教育水平和公共服务能力低，一方面使得人口素质低，发展生产的能力不强，自身难以脱贫致富，另一方面社会保障水平低，困难群众难享改革发展的红利。四是产业结构单一，发展能力薄弱，产业化落后；五是扶贫资金明显不足。我们针对以上问题，提出对口解决方案。

3 模型假设

1. 假设异常值处理后的数据正确且有效，能够正确反映各个村庄的帮扶绩效。

2. 假设脱贫帮扶绩效评价机制中居民收入、产业发展、居住环境、文化教育和基础设施五个评价指标均为主要影响因素，与绩效成线性相关关系。

3. 假设方差分析中各个总体服从正态分布且各总体的方差相同。

4 符号表示

表 1 变量符号及其解释说明

符号	解释说明
ΔY	总分相对增长量（2020年扶贫总分－2015年扶贫总分）
X_1	居民收入相对增长量指标
X_2	产业发展相对增长量指标
X_3	居住环境相对增长量指标
X_4	文化教育相对增长量指标
X_5	基础设施相对增长量指标
$F_i\left(i=1,2,3,4,5\right)$	因子得分
Y	是否被评为脱贫先进村庄（可评为=1，不可评为=0）
auc	二元 Logistic 预测准确率

5 问题一

5.1 数据预处理

由于测量数据在其采集与传输过程中，由于环境干扰或人为因素有可能造成个别数据不切合实际或丢失，即出现异常值。为了恢复数据的客观真实性以便将来得到更好的分析结果, 有必要先对原始数据剔除异常值。

另外，无论是人工观测的数据还是由数据采集系统获取的数据，都不可避免叠加上“噪声”干扰，反映在曲线图形上就是一些“毛刺和尖峰”。为了提高数据的质量，必须对数据进行平滑处理去噪声干扰。

5.1.1 处理异常值

1. 基本思想

规定一个置信水平，确定一个置信限度，凡是超过该限度的误差，就认为它是异常值，从而予以剔除。

2. 基本原理

假设数据依正态分布，我们使用拉伊达方法（非等置信概率）剔除异常值。如果某测量值与平均值只差大于标准偏差的三倍，则被视为异常值，应为所给数据均已被标准

化，所以我们采用以 0 来代替异常的方法来修正。

其中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 为样本均值，

$S_x=\sqrt{\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)}$ 为样本的标准偏差。

5.1.2 数据的平滑处理—“加权移动平均”平滑滤波

1. 加权基本思想

作平均的区间内中心处数据的权值最大，愈远离中心处的数据权值越小。这样就减少了对真实信号本身的平滑作用。权重系数可以采用最小二乘原理，使平滑后的数据以最小均方差逼近原始数据。即令：

$$\min\sum_{k}\left(y'_{i+k}-y_{i+k}\right)^2 \quad (1)$$

我们采用“五点二次平滑”方法（n=5,k=-2,-1,0,1,2），利用如下公式（2）进行计算

$$\begin{cases}\sum_{k=-2}^{2}\left(y_{i+k}-A_0-A_1k-A_2k^2\right)=0\\ \sum_{k=-2}^{2}\left(y_{i+k}-A_0-A_1k-A_2k^2\right)k=0\\ \sum_{k=-2}^{2}\left(y_{i+k}-A_0-A_1k-A_2k^2\right)k^2=0\end{cases}$$

2. 处理结果

利用 MATLAB 软件，结合上述原理对本文中的数据进行处理。表 2 为五点二次平滑权重系数结果，图 2 中以 2015 年的 SR、CY 两组数据为例，展示了平滑前后的对比图。从图中我们可以看出，修正后的数据分布更加集中，这对本文后续的建模分析有非常重要的作用，修正后的数据见支撑材料。

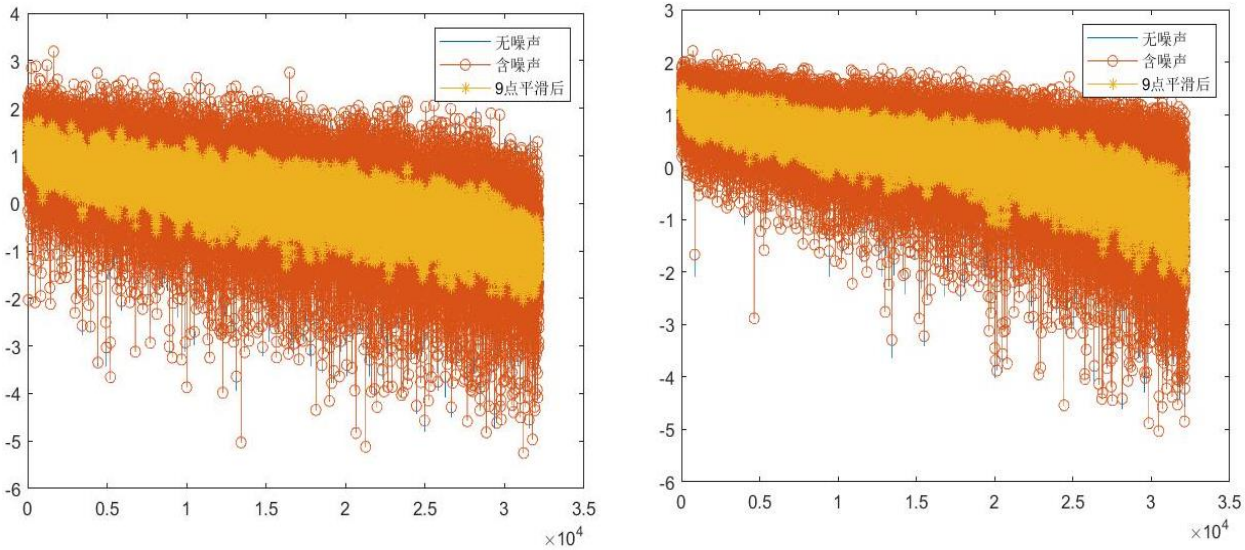


图 2 数据平滑前后对比图

表 2 五点二次平滑权重系数表

	归一系数	y_{-2}	y_{-1}	y_0	y_1	y_2
y'_{-2}	35	31	9	-3	-5	3
y'_{-1}	35	9	13	12	6	-5
y'_0	35	-3	12	17	12	-3
y'_1	35	-5	6	12	13	9
y'_2	35	3	-5	-3	9	31

5.2 相关性分析

5.1.1 方法概要

为了证实问题一中所说的假设—五年前的评分与 2020 年对应的各项评分有着直接的关联，我们采用相关性分析的方法来验证。所谓相关性分析，就是指研究现象之间是或否存在某种依存关系，并对具体有依存关系的现象探讨其相关程度，是研究随机变量之间的相关关系的一种统计方法。简单地说，相关分析就是衡量两个数值型变量的相关性，以及计算相关程度的大小[2]。

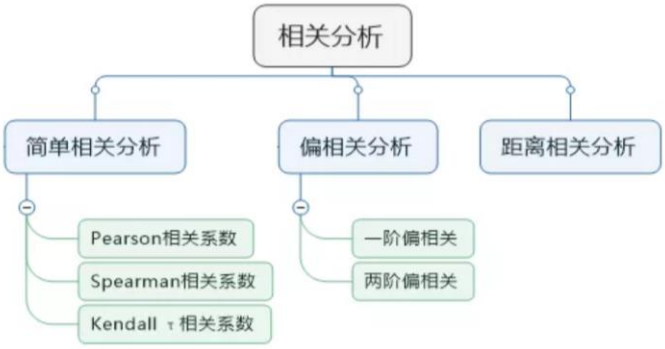


图 3 相关分析种类

图 3 展示了所有的相关性分析的种类，对于本问题来说，我们进行简单的相关分析就可以。线性相关分析：研究两个变量间线性关系的程度。用相关系数 r 来描述。表 3 展示了 r 与两变量之间的相关程度之间的关系。

表 3 相关系数 r 与相关程度关系

相关系数	相关程度
$0.0 \le \lvert r \rvert \le 0.3$	不相关
$0.3 \le \lvert r \rvert \le 0.5$	弱相关
$0.5 \le \lvert r \rvert \le 0.8$	中度相关
$0.8 \le \lvert r \rvert \le 1.0$	强相关

关于 r 的计算有三种：①Pearson 相关系数：对定距连续变量的数据进行计算。②Spearman 和 Kendall 相关系数：对分类变量的数据或变量值的分布明显非正态或分布不明时，计算时先对离散数据进行排序或对定距变量值排（求）秩。

在进行简单相关分析时，首先需要绘制散点图来判断变量之间的关系形态，如果是线性关系，则可以用相关系数来测度两个变量之间的关系强度，然后对相关系数进行显著性检验，以判断样本所反映的关系能否代表两个变量总体上的关系。图 4 展示了简单

相关分析的基本步骤。

图 4 简单相关分析步骤

5.1.2 模型建立与求解

1. 绘制散点图，观察两变量是否有规律变化。

对附件中所给的得到帮扶村庄的居民收入(记为 SR)、产业发展(记为 CY)、 居住环境(记为 HJ)、文化教育(记为 WJ)、基础设施(记为 SS)等五个评价指标的数据，按照时间两两配对(即 2015 年与 2020 年 SR 配对、2015 年与 2020 年 CY 配对等)分成 5 组。利用 SPSS 软件来绘制 5 组数据的散点图。

以居民收入(SR)这组数据为例，散点图如图 4a)所示，从图中能看出两组数据呈现简单的线性关系。同时，将其他几组变量的散点图也做出来，结果如图 4b)散点图阵所示。从图中我们可以看出，5 组变量之间均存在一定的相关关系，所以我们可以对其进行下一步分析。

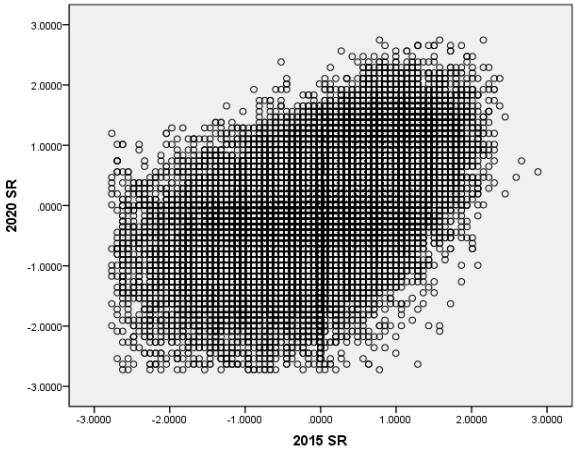


图5a) 2015年与2020年SR散点图

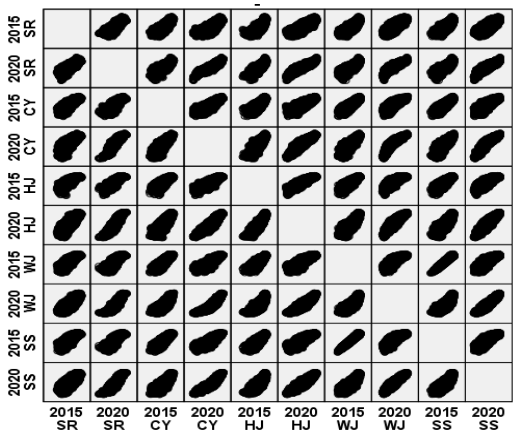


图5b) 各组变量之间的散点图阵

图 5 散点图

2.根据变量类型或正态性检验，选择合适的相关系数公式。

本文所采用的数据可以看为连续的数值型数据，对于相关系数的选择主要是判断其是否满足正态分布。利用 SPSS 软件中检验单样本正态分布的 K-S 检验来判断这些变量是否服从正态性分布，部分检验结果如表 4 所示。从中我们可以看出这 10 个变量检验的 p 值均小于显著性水平 0.05，说明其不满足正态分布假设。所以，我们将采用非参数方法中的 Spearman 相关系数来研究 5 组数据之间的相关程度。

表 4 正态性检验(单样本 Kolmogorov-Smirnov 检验)

	2015 SR	2015 CY	2015 HJ	2015 WJ	2015 SS	2020 SR	2020 CY	2020 HJ	2020 WJ	2020 SS
N	32155	32155	32155	32155	32155	32155	32155	32155	32155	32155
均值	.08877	.27002	.19342	.26544	.26889	.05269	.06072	.00712	.00455	.02165
标准差	.8910350	.7050574	.7173227	.7779923	.7210800	.8947760	.9017319	.9734912	.9802052	.9777902
统计量值	10.378	16.557	19.563	15.668	19.253	8.410	10.699	9.572	5.631	5.378
P 值(双侧)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

3.计算 Spearman 秩和相关系数 r 来评价相关程度。并进行显著性检验，如果 $P<\alpha$，

表示存在显著相关性。检验的假设为：两变量之间无显著行线性关系，即两个变量之间存在零相关，检验结果如表 5 所示。

根据表 5 中所显示的结果来看，在显著性水平为 0.01 时，检验的 p 值均小于 0.01，说明通过显著性检验，即 5 组变量每组中的两个变量之间有显著线性关系。且从相关系数来看，r 的取值满足 $0.5 \le r \le 0.8$，说明变量之间存在中度相关。即问题一中的“，五年前的评分与 2020 年对应的各项评分有着直接的关联”观点成立，且各个评价指标之间均存在正的线性相关关系。

表 5 Spearman 相关系数显著性检验

		2015 SR	2015 CY	2015 HJ	2015 WJ	2015 SS
2020 SR	相关系数	.526	.484	.568	.511	.518
	Sig.（双侧）	.000	.000	.000	.000	.000
2020 CY	相关系数	.484	.637	.588	.644	.647
	Sig.（双侧）	.000	.000	.000	.000	.000
2020 HJ	相关系数	.582	.587	.786	.610	.620
	Sig.（双侧）	.000	.000	.000	.000	.000
2020 WJ	相关系数	.449	.584	.542	.664	.627
	Sig.（双侧）	.000	.000	.000	.000	.000
2020 SS	相关系数	.457	.558	.553	.632	.628
	Sig.（双侧）	.000	.000	.000	.000	.000

6 问题二

6.1 帮扶类型编号排序

6.1.1 数据说明

运用对数据进行平滑处理以对异常值修正后的数据作为原始数据，以编号自 0 至 5 的指定帮扶类型作为研究对象。以 2020 年标准化后的总分指标值减去 2015 年的总值指标作为扶贫评价指标：总值增长相对量，记作 ΔY

$$\Delta Y = 2020\text{总分} - 2015\text{总分} \quad (3)$$

利用数据库处理对数据按照 6 个帮扶类型进行分类，并求出每组内总分增长相对量的平均值，作为评价指标体系，处理后数据见表 6。

表 6 数据处理结果表

帮扶单位类型	0	1	2	3	4	5	总计
N	2483	12492	2131	11279	29	3741	32155
均值	-.159015	-.242129	-.183181	-.245324	-.675006	-.111019	-.218062

6.1.2 非参数方法：Jonckheere-Terpstra 检验

1. 方法概要

Jonckheere-Terpstra 检验在多个独立样本非参数检验中，可以检验多个独立样本的位置参数是否持续上升和下降（备择假设具有方向性）。在检验多个独立样本，比如检验几种不同方法、决策或者实验条件所产生的结果是否相同时，如果数据不满足正态分

布或者数据的测度水平是定序水平的，那么就可以使用非参数统计方法，多个独立样本非参数统计中常用的方法有 Kruskal-Wallis 检验和中位数检验。

假定 k 个独立样本(总体)分别来自有同样形状的连续分布函数，各其位置参数(比如中位数)记为θ_1，$\theta_2,\cdots,\theta_k$。假定 k 个样本(总体)的样本量为$n_i, i=1,2,\cdots,k$。令x_{ij}为来自第i个样本(总体)的第 j 个独立观测值。

那么，观测值x_{ij}可以写成下面的线性模型：（误差独立同分布）

$$x_{ij}=\mu+\theta_i+\varepsilon_{ij}, i=1,2,\cdots,k, j=1,2,\cdots,n_i \quad (4)$$

2. 模型建立及求解

Jonckheere-Terpstra 检验步骤：①设立假设检验问题，因为本题中我们采用位置参数是各个样本的均值，并要求对分类型自变量进行排序，所以建立如下单侧检验的假设检验问题：

$$\begin{aligned} H_0&:\mu_0=\mu_1=\mu_2=\mu_3=\mu_4=\mu_5 \\ H_1&:\mu_0\le\mu_1\le\mu_2\le\mu_3\le\mu_4\le\mu_5 \end{aligned} \quad (5)$$

②计算统计量。对所有的U_{ij}在$i<j$范围求和，这样就产生了 Jonckheere-Terpstra 统计量：

$$V=\sum_{i<j}U_{ij} \quad (6)$$

③做决策。在给定水平α，通过查表可以得到在零假设条件下的临界值，当 J 较大时，应拒绝原假设。当样本量较大，可以利用正态近似，这时，在给定水平α下，如果$J=E_{H_0}(J)+Z_\alpha\sqrt{\mathrm{var}_{H_0}(J)}$，拒绝零假设。

依据以上检验步骤，结合 SPSS 软件操作，最终得到的检验结果如下图 6 所示。从图中我们可以看出，在显著性水平 0.05 水平下，做出的决策为：拒绝原假设，即认为 6 种不同的扶贫类型之间存在显著性差异。

假设检验汇总

	原假设	测试	Sig.	决策者
1	相对增长量 的分布在 帮扶单位类型 (0-5) 类别上相同。	排序替代选项的独立样本 Jonckheere-Terpstra 检验	.046	拒绝原假设。

显示渐进显著性。显著性水平是 .05。

图 6　J-T 检验结果

6.1.3 参数方法：方差分析

1.方法概要

方差分析是检验多个总体均值是否相等的统计方法，但本质上它所研究的是分类型自变量对数值型因变量的影响，换句话说就是，通过检验各总体的均值是否相等来判断分类型自变量对数据型因变量是否有显著影响。

当方差分析中只涉及一个分类型自变量时，称为单因素方差分析。在单因素方差分析中，用 A 表示因素，因素的 k 个水平（总体）分别用$A_1,A_2,\cdots,A_k$表示，每个观测值

用 $x_{ij}\left(i=1,2,\cdots,k;\ j=1,2,\cdots,n\right)$ 表示，即 x_{ij} 表示第 i 个水平（总体）的第 j 个观测值。其中，从不同水平中所抽取的样本量可以相等，也可以不相等。

2. 模型建立

在做方差分析时需要满足数据服从正态分布，所以我们假设每个总体都服从正态分布，且各总体的方差相同。在此假设下，做进一步分析。基本步骤如下：

①提出假设。与非参检验假设相同。

②构造检验统计量。

$$F=\frac{MSA}{MSE}\sim F\left(k-1,n-k\right)\quad（7）$$

其中，$MSA=\dfrac{SSA}{k-1}=\dfrac{\sum_{i=1}^{k}n_i\left(\overline{x_i}-\overline{\overline{x_i}}\right)^2}{k-1}$，

$MSE=\dfrac{SSE}{n-k}=\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2}{n-k}$，

n 为全部观测值的个数，k 为因素水平（总体）的个数。

③做出决策。根据给定的显著性水平 α，在 F 分布表中查找与分子自由度 $df_1=k-1$、分母自由度 $df_2=n-k$ 相应的临界值 $F_\alpha\left(k-1,n-k\right)$。若 $F>F_\alpha$，则拒绝原假设，表明 μ_i 之间存在显著性差异；反之，不能拒绝原假设。对于 p 值，若 $P<\alpha$，这拒绝 H_0。

3. 模型求解

依据以上检验步骤，结合 SPSS 软件操作，最终得到的检验结果如下表 7 所示。从表中的结果数据可以看出，在显著性水平 0.05 水平下，p 值均小于 α=0.05，所以做出的决策为：拒绝原假设，即认为 6 种不同的扶贫类型之间存在显著性差异。这个结果与非参数检验结果相互证实。

表 7 方差分析结果表

	平方和	df	均方	F	显著性
组间组合	75.788	5	15.158	40.561	.000
组间未加权线性项	3.825	1	3.825	10.237	.001
组间加权项	11.476	1	11.476	30.710	.000
组间偏差	64.312	4	16.078	43.024	.000
组内	12014.066	32149	.374		
总数	12089.854	32154			

6.1.4 排序结果

有两种检验结果来看，6 种类型的帮扶单位的均值之间存在着显著性差异。根据变量的意义来看，均值高的表示该帮扶单位有较高的绩效。所以我们可以按照总分相对变量增长量 $\left(\Delta Y\right)$ 的均值大小来对这 6 种类型的帮扶单位进行排序。

由图 7 可以得出最终的排序结果，6 种类型的帮扶单位的绩效由高到低依次为：5、0、2、1、3、4。

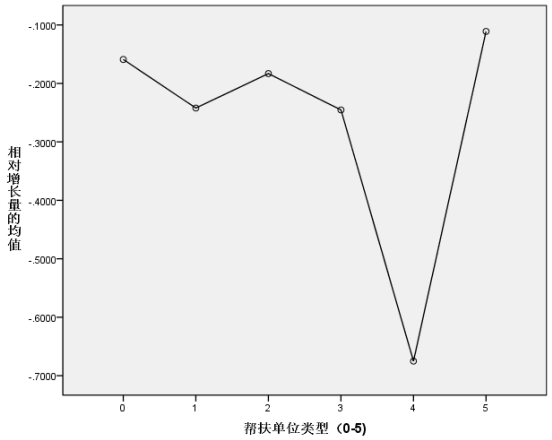


图 7 相对增长量均指图

6.2 帮扶单位编号排序—T0PSIS 模型

6.2.1 数据说明

运用对数据进行平滑处理以对异常值修正后的数据作为原始数据，以编号自 0 至 159 个指定帮扶单位作为研究对象。以 2020 年五个标准化后指标值减去 2015 年相对应的值作为扶贫评价指标：增长相对量，利用数据库处理对数据按照 160 个帮扶单位进行分类，并求增长相对量的平均值，作为评价指标体系，部分处理后数据见表 8[3]。

表 8 数据处理结果表

帮扶单位编号	居民收入相对增量	产业发展相对增量	居民环境相对增量	文化教育相对增量	基础设施相对增量
0	-0.34	-0.01	0.32	0.06	0.25
1	0.04	-0.01	0.02	-0.08	-0.06
2	0.01	-0.10	-0.09	-0.19	-0.16
3	-0.13	-0.27	-0.22	-0.57	-0.39
4	-0.24	-0.45	-0.26	-0.48	-0.62
5	-0.25	-0.39	-0.15	-0.57	-0.59
6	-0.04	-0.21	-0.23	-0.26	-0.27
7	0.00	-0.21	0.17	-0.36	-0.05
8	-0.24	-0.28	-0.37	-0.30	-0.35
9	-0.07	-0.24	-0.12	-0.40	-0.24
10	-0.15	-0.22	-0.15	-0.34	-0.43

6.2.2 模型概要

TOPSIS 法（Technique for order preference by similarity to ideal solution），是有限方案多目标决策分析的一种常用方法。其用于研究与理想方案相似性的顺序选优技术，通俗理解即为数据大小有优劣关系，数据越大越优，数据越小越劣，结合数据间的大小，计算正负理想解和正负理想解距离与综合得分，最后得出优劣方案的排序。TOPSIS 分析通常包括以下五步：①指标数据的属性趋同化处理；

②对趋同化的数据归一化处理，并得到归一化处理后的矩阵 Z；

③找出最优和最劣矩阵向量；

④分别计算评价对象与正理想解距离 D^+或负理想解距离 D^-；

$$D_i^+ = \sqrt{\sum_{i=1}^{m}(\max Z_{ij} - Z_{ij})^2} \qquad D_i^- = \sqrt{\sum_{i=1}^{m}(\min Z_{ij} - Z_{ij})^2} \quad (8)$$

⑤计算各评价对象与最优方案的接近程度 C_i 值，并且进行排序，得出结论。

6.2.3 数据预处理：

第一步：首先数据一定需要全部同趋势正向化（所有的数据表示为数字越大，评分越大）。由于题干中所提供数据均满足此要求，因此不需要再进一步处理

第二步：对处理后的数据再次进行归一化处理并求平方。

6.2.4 模型建立及求解

针对 5 个指标进行 T0PSIS 评价，其评价对象为 160 个（帮扶单位编号 0-159）。合理的确定指标的权重是应用 TOPSIS 综合评价的关键，由于无法获取专家打分，本文不选取主观性的权重来建立模型。选择利用基于信息论的熵值法，根据各指标所含信息的有序程度的差异性来确定指标权重的客观赋权方法，仅依赖于数据本身的离散程度，如果指标的离散性程度越大，那么熵值越少，这表明该指标提供的信息越多，因此应当对该指标取得较大的权重。

利用 python 中的 pandas、statsmodels 和 scipy 模块计算熵值和各指标的权重，得出这 5 项指标的权重分别为：1.2897511，-0.09997849，-0.06776056，-0.05919084，-0.06282122。利用熵增法确定权重后，计算这 5 个评价指标的正负理想解值，接着计算帮扶单位评价对象与正负理想解的距离值 D^+和 D^-。根据 D^+和 D^-值，最终计算得出综合得分指数，并可针对其值进行排序。

表 9 正负理想解计算结果表

	x1	x2	x3	x4	x5
负理想解	-0.72	-1.2	-1.68	-0.89	-0.93
正理想解	0.63	0.6	0.32	0.26	0.4

Topsis 评价计算结果见表 10，对给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号分别为：142，60，74，137，144，112，158，76，147，114。并作出这十个单位的五类指标分布散点图，见图 8。

表 10 排序结果表

编号	x1	x2	x3	x4	x5	正理想解	负理想解	综合得分指数	排序
142	0.22	-0.38	-0.21	-0.58	-0.45	0.075	0.95	0.926775	1
60	0.08	-0.48	-1.10	-0.84	-0.46	0.10	0.86	0.896725	2
74	0.28	-0.24	-0.33	-0.08	-0.06	0.18	0.99	0.844166	3
137	0.28	-0.07	-0.15	-0.22	-0.19	0.24	0.97	0.803584	4
144	0.15	-0.52	-0.39	-0.55	-0.4	0.22	0.89	0.799103	5
112	0.30	0.14	-0.17	-0.13	-0.15	0.26	0.97	0.786302	6
158	0.23	-0.32	-0.28	0.01	-0.1	0.27	0.93	0.775227	7
76	0.01	-0.76	-1.04	-0.49	-0.66	0.27	0.79	0.748632	8
147	0.21	-0.13	-0.27	-0.35	-0.26	0.31	0.91	0.746941	9
114	0.3	0.21	-0.17	0.2	0.16	0.32	0.92	0.741685	10

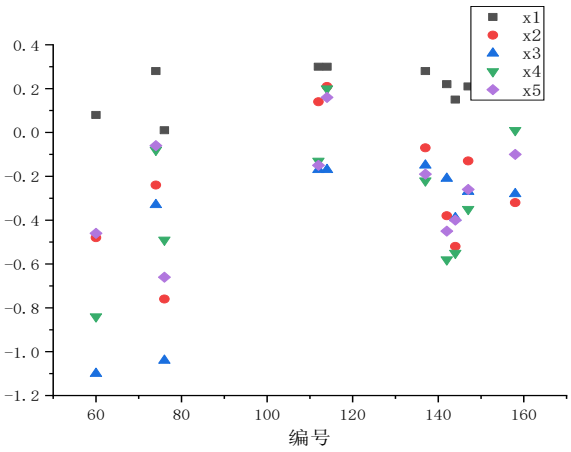


图 8 前十名帮扶单位指标增长量散点图

7 问题三

7.1 数据选择与分析

由于每个帮扶单位在扶贫上有着不同的工作特色，在对帮扶单位的总体绩效进行排序后仍然有必要对居民收入、产业发展、居住环境、文化教育、基础建设这 5 类评价指标上分别进行研究，以确定帮扶业绩明显的帮扶单位[4]。

由此，在第二问的数据汇总基础上，显然每一项指标只有相对增量是远远不够的，为此加入这 5 个指标的 2015 年和 2020 年的原始值（x6-x15）和总分相对增量 x16（2020 年标准化评价总分-2015 年标准化评价总分），一共 16 个指标衡量不同帮扶单位在扶贫上的绩效，利用因子分析法，对因子进行旋转后指定提取 5 个因子并对因子的经济含义进行解释，最后计算每一个帮扶单位在 5 个因子上的因子得分，分别进行排序得出结果。16 个指标的部分帮扶单位的数据见表 11。

表11 问题三部分数据表

单位编号	x1	x2	x3	x4	x5	x6	x7	x8
0	-0.34	-0.01	0.32	0.06	0.25	1.06	1.06	0.90
1	0.04	-0.01	0.02	-0.08	-0.06	0.85	0.92	0.82
2	0.01	-0.10	-0.09	-0.19	-0.16	0.61	0.77	0.67
3	-0.13	-0.27	-0.22	-0.57	-0.39	0.44	0.66	0.55
4	-0.24	-0.45	-0.26	-0.48	-0.62	0.27	0.50	0.38
5	-0.25	-0.39	-0.15	-0.57	-0.59	0.30	0.53	0.41
单位编号	x9	x10	x11	x12	x13	x14	x15	x16
0	1.02	1.03	1.02	1.17	1.23	1.34	1.30	0.17
1	0.94	0.96	0.84	0.97	1.02	1.07	1.02	0.06
2	0.80	0.82	0.50	0.67	0.68	0.67	0.64	-0.14
3	0.69	0.71	0.33	0.45	0.46	0.39	0.38	-0.25
4	0.50	0.51	0.06	0.15	0.13	0.02	0.00	-0.45
5	0.53	0.53	0.12	0.20	0.20	0.06	0.04	-0.44

7.2 建立模型

7.2.1 适用性检验

首先考察这 116 个指标是否存在一定的线性关系，是否适合采用因子分析和提取因子。利用 SPSS16.0 软件，计算变量的相关系数矩阵，得出大部分相关系数都较高(大于 0.3)，通过检验。

利用巴特利特球形度检验和 KMO 检验方法进行分析，结果显示：巴特利特球形度检验统计量的观测值为 8001.251，相应的概率 p 值接近 0，如果显著性水平 a 为 0.05，由于概率 p 值小于显著性水平 a，则应拒绝原假设，认为相关系数矩阵与单位矩阵有显著差异。同时，KMO 值为 0.901，根据 KMO 度量标准可得原有变量比较适合进行因子分析[5]。

7.2.2 因子旋转与命名

在假设条件下，利用主成分方法求解因子载荷矩阵，根据原有变量的相关系数矩阵，指定选取 5 个特征值，为了对因子进行命名与更好的分析，利用方差最大化方法正交旋转因子矩阵，使得因子载荷要不和 1 接近要不和 0 接近，得出因子解释原有变量总方差的情况以及旋转后的成分矩阵，见表 12[6]。

表12 旋转后的成分矩阵

	1	2	3	4	5
x1	-.204	.278	.932	.021	.083
x2	.354	.591	.241	.115	.656
x3	.688	.577	.072	.109	-.112
x4	.123	.919	.176	-.023	.077
x5	.160	.891	.072	.197	.102
x6	.973	.129	-.113	-.005	.077
x7	.981	.129	-.096	.009	.080
x8	.971	.154	-.029	.099	.090
x9	.986	.106	-.082	.043	.063
x10	.986	.104	-.081	.039	.064
x11	.952	.205	-.012	.202	.074
x12	.937	.219	-.021	.214	.117
x13	.947	.207	-.024	.217	.079
x14	.926	.227	-.059	.255	.099
x15	.921	.245	-.055	.274	.091
x16	.620	.396	.068	.657	.111

由表 12 可知，x3 在第 1 个因子上有较高的载荷，即第 1 个因子主要解释了这几个变量，可将解释为以居住环境基为代表的增长因子。x4、x5、x6、x15 在第 2 个因子上有较高的载荷，第 2 个因子主要解释了这几个变量，可解释为文化教育因子。第三个因子解释了 x1，可解释为居民收入因子。第四个因子解释了 x16，可解释为基础建设因子。第五个因子解释了 x2，可相对解释为含有产业发展因子。与旋转前相比，因子含义较清晰。由此试图从旋转后的矩阵对因子再次进行对 5 个指标进行含义说明。并作出因子旋转组件图，见图 9。

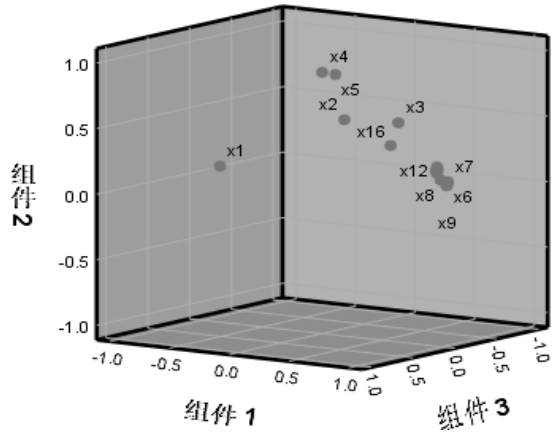


图9 因子旋转组件图

7.2.3 计算因子得分

因子模型建立以后，为了建立综合评价体系，需要估计因子得分系数，并输出因子得分系数，并计算因子得分系数，以建立评价体系并排序。可以利用原始的因子载荷矩阵换为主成分分析，根据本问题要求，对每一个因子得分进行排序即可，并列出每一个因子排名前 5 的企业类型与得分，结果见表 13。

表13 排序结果表

企业编号	F1（居住环境）	企业编号	F2（文化教育）	企业编号	F3（居民收入）	企业编号	F4（基础建设）	企业编号	F5（产业发展）
1	1.80785	149	2.9643	115	2.74619	106	2.24036	72	3.90976
0	1.75341	156	2.66788	54	2.31486	108	1.72758	66	3.33065
21	1.70919	155	2.52861	134	2.13287	113	1.62806	120	2.58017
25	1.64477	41	2.52012	131	1.75838	128	1.62396	79	1.84224
7	1.63415	130	1.9279	142	1.7498	110	1.61421	85	1.51492

在因子分析下，帮扶单位 1，0，21，25 和 7 可能在改善居住环境上帮助的效果不错。帮扶单位 149，156，155，41 和 130 可能在改善文化教育上帮助的效果不错。帮扶单位 115，54，134，131 和 142 可能在提高居住收入上帮助的效果不错。帮扶单位 106，108，113，128 和 110 可能在改善基础建设上帮助的效果不错。帮扶单位 72，66，120，79 和 86 可能在提高产业帮助上的效果不错。

值得注意的是，因子模型适应性效果很好，提取的因子能较好的符合实际且有明确的解释效果和信息囊括能力，基于假设下，构建的评价指标体系，比较符合实际合理。但受到数据选择的限制，因子的命名与解释需要改进。在本文中因子的命名是以不相似以及和题目相对应进行的命名，实际上 F1 不单单只是解释了居住环境指标，其已经囊括了其余类型的大部分，其余因子含义均是如此。因此，为了更好的对这 5 类指标进行排序，还需要在此基础上进行进一步的提炼与改进。

8 问题四

8.1 二元 Logistic 回归

8.1.1 数据处理与分析

原始数据共计 32155 个数据，要选其中 1 万个为先进单位，先进单位占总体的比例

约为 31%。利用第二问 Topsis 综合评价法下综合排名的 160 个数据结果，取前 41 名评为先进单位，记为 1，其余为 0。将是否评选为先进单位 y(1 和 0)为因变量，其余五个指标的增长相对值为自变量，做非线性回归：二元 logistics，探究这些因素对评选先进单位的影响[7]。

8.1.2 模型概要

Logistic，是一种广义的线性回归分析模型，主要思想是，根据现有数据对决策边界建立回归方程，然后将回归方程映射到分类函数上实现分类。其适用条件：因变量为二分类的分类变量或某事件的发生率，并且是数值型变量；残差和因变量都要服从二项分布；自变量和 Logistic 概率是线性关系；各观测对象间相互独立[8]。

Logistic 回归的原理可以理解为以下四步：

1、利用回归方程表示决策边界。分类问题的目的是找到决策边界，因此我们需要找到一个回归方程来表示这个决策边界： g(W,X)=W^X ，其中 W 代表权重向量；

2、利用 Sigmoid 函数对回归关系进行映射；在面对二分分类问题时，可以用 1 和 0 分别代表一种情况，此时利用 Sigmoid 函数；sigmoid 函数图像见图 10。

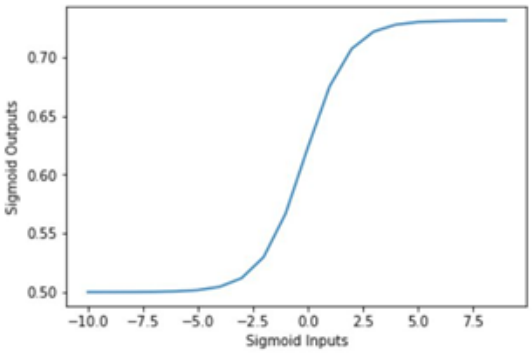


图 10 sigmoid 函数

3、在得到拟合函数后，利用损失函数来评价模型与实际值之间的差异大小；

4、求出损失函数取得极小值时对应的 W ，从而得到拟合函数[9]。

Logistic 回归的评价指标：

1、混淆矩阵(confusion matrix)

包括分类器预测结果：真正 TP(true positive)、真负 TN(true negative)、假正 FP(false positive)、假负 FN(false negative)的数量，其中真正和假负均为正确分类的结果。

2、准确率、真正率及假正率

预测误差(error)和准确率(accuracy)都可以表示误分类样本数量的相关信息，真正率(TPR)和假正率(FPR)也是很有参考价值的性能指标， TPR={TP}/{AP} 表示预测与实际均为正类别样本数量 与 实际正样本数量的比值，FPR={FP}/{AN} 表示预测为正类别实际为负类别样本数量 与 实际负样本数量的比值。

3、ROC 曲线(receiver operator characteristic)

ROC 曲线由变量 1-Specificity 和 Sensitivity 绘制，其中横轴为假正率(FPR)、纵轴为真正率(TPR)，ROC 曲线的对角线表示随机猜测，若 ROC 曲线在对角线下表示分类器性能比随机猜测还差，ROC 曲线下的区域面积(AUC)表示分类模型的性能，反映了模型将正例排在反例前的比例（当 AUC=1 时，说明将所有正例均排在反例之前）[10]。

8.1.3 影响因素分析

利用 python 软件建立模型，利用 LogisticRegression 函数模型，并使得 C(正则化系数)较大，以使得结果与利用 statsmodel 后参数结果较接近，并输出截距与斜率值。回归

结果见下式：

$$p=\frac{\exp(-9.793+66.067x_1\text{-}6.958x_2-5.793x_3\text{-}1.826x_4\text{-}4.756x_5)}{1+\exp(-9.793+66.067x_1\text{-}6.958x_2-5.793x_3\text{-}1.826x_4\text{-}4.756x_5)} \quad (9)$$

将结果与熵值法所计算出的指标权重进行综合分析显示，居民收入相对增长量指标对脱贫先进村庄评级影响力度最大，具有重要影响，且成正相关。其次是产业发展相对增长量指标，居住环境相对增长量指标和基础设施相对增长量指标，最后是文化教育相对增长量指标。尽管这些呈现负相关，但是由于指标是相对性标准化后，对其系数意义解释没有充分性含义[11]。

8.1.4 模型评价

根据模型结果进行评价，得出混淆矩阵为：

$$\begin{pmatrix} 92 & 1 \\ 43 & 7 \end{pmatrix}$$

并画出 ROC 曲线，靠近左上角的 ROC 曲线代表的学习器准确性越高，同时 AUC 考虑了学习器对于正例和负例的分类能力，在样本不平衡的情况下，依然能对分类器做出合理评价。混淆矩阵热力对应图和 ROC 曲线见图 11。

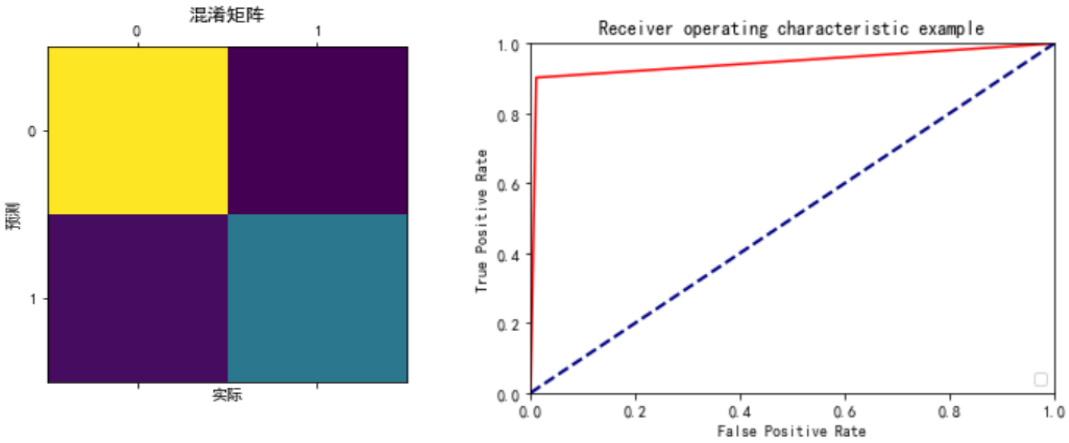


图 11 混淆矩阵热力对应图及 ROC 曲线

由表 14 得出，模型预测准确率（auc）达到 0.946，召回率（call）达到 0.825，模型适用性良好。

表 14 模型评价结果表

	precision	recall	f1-score	support
0	0.96	0.99	0.97	93
1	0.97	0.9	0.94	41
accuracy			0.96	134
macro avg	0.97	0.95	0.96	134
weighted avg	0.96	0.96	0.96	134

8.2 数据补充

为了能够解决接下来村庄编号的排序及级别评选，必须将 2020 年的相应数据补充出来，一般会选用参数估计与类比估算的方法。基于本数据集样本量大，样本数据充足

特点，我们采取类比估计法对未知数据进行补充。类比估算法也被称作自上而下的估算，是一种通过比照已完成的类似项目的实际成本，去估算出新项目成本的方法。类比估算法法师和品谷一些与历史项目在应用领域、环境和复杂度方面相似的项目。以村庄编号为 39257 的 2020 年 SR 数据为例，介绍本文所采用的类比估计法的原理及操作步骤。

①首先，可以看到该村庄的帮扶单位及帮类型分别为 116、1，所以在对修正后的原始数据按照编号及单位排序后，筛选出对应的帮扶单位及类型的所有村庄编号。

②在筛选出来的村庄编号中，找到与编号 39257 对应的 2015 年 SR 数值(即 1.64900)相同或相近的村庄编号（若是相同的村庄编号较少，则可以进一步选出与其临近的村庄编号），将其筛选出来。

③计算出满足要求的村庄编号的 SR 相对增长值，求其平均值，则编号为 39257 的 2020SR=2015$SR+\overline{X_1}$ 2020SR。

④若与 2015 年 SR 数值相近的村庄编号较多(大于等于 10)，可以直接用用相应的 2020 年 SR 均值来补充编号为 39257 的 2020 年 SR。

参照上述方法及操作步骤，对其余村庄编号的 2020 年各变量数据进行补充，得到的结果如表 15 所示，根据结果可以做进一步的研究。

表 15 数据填补结果

村庄编号	帮扶单位 (0-159)	帮扶单位类型（0-5)	2020 SR	2020 CY	2020 HJ	2020 WJ	2020 SS
39257	116	1	1.321764	0.894708	1.009125	1.120231	0.344212
25149	89	1	1.287797	0.8457156	1.238553	1.301921	1.27345
12722	7	2	0.588924	0.575556	0.873498	1.050967	0.935013
12916	10	2	0.61313	0.327901	0.72478	0.539768	0.989891
21570	47	1	-0.312528	0.129989	0.834072	0.34658	0.577488
22096	48	1	0.050657	0.365088	-0.167233	0.101639	-0.223065
47883	138	1	0.874913	0.374536	1.21329	0.522161	0.713174
34208	78	1	0.693456	0.518277	0.632014	0.067594	0.134734
34276	78	1	0.296457	1.031275	0.782247	0.568312	0.364028
52436	151	3	-0.40277	-0.386882	-0.02508	0.16838	-0.39642

8.3 级别评价

利用上文已补充的数据，再次计算相对增长量，作为模型中的 x1 至 x5，并带入 Logistic 回归的方程中,计算 y 值[13]，并对其进行排序以划分级别。其计算结果见表 16。

表 16 模型计算结果与评级表

村庄编号	SR 相对增量（x1）	CY 相对增量（x2）	HJ 相对增量（x3）	WJ 相对增量（x4）	SS 相对增量（x5）	$\hat{y}$
22096	0.067267	-0.64441	-0.17759	-0.98406	-0.61663	0.992559195803657
34208	-0.01412	0.010697	-0.23267	0.097524	-0.25883	0.000224855373999
52436	-0.09649	-0.21007	-0.29174	-0.47107	-0.03201	0.000006121724159
47883	-0.04993	0.095086	0.30589	0.239711	0.130124	0.000000062898825

可以看出，村庄编号为 22096 被评为先进村庄的概率接近于 1，表明其有极大的概率能够被评为先进村庄。而其余的九个数据的概率均较低，被评为先进村庄的概率约为 31%，且其中和其他概率值相比，具有显著性数量级差异的是村庄编号为 34208，52436 和 47883，表明其仍然存在着一定的概率。因此，不仅可以将 22096 评价为贫困先进村庄，根据需要也可以增添 34208，52436 和 47883 村庄。若称号分为一级和二级，且一二级的称号比例为 1:3，那么就可以将 22096 评为一级村庄。34208，52436 和 47883 称为二级村庄。

9 问题五

尊敬的国家扶贫办领导：

您好！

根据我们团队的研究成果，贫困地区的现状和存在的主要问题主要表现为贫困人口多、基础设施薄弱、受教育水平和农村公共服务能力低、产业结构单一和发展能力薄弱、扶贫资金不足等五个方面。针对这些问题，本团队提出如下对策和建议。

一、加强对口帮扶责任制

加强扶贫力度，严格实行目标责任考核。帮扶的部门要派工作队，长期驻在扶贫点上，帮助被扶持的地区发展生产，改善生产生活环境，帮助搞好基层组织精神文明建设;与此同时,从物力、财力等方面给予全力支持，加大对扶贫开发对口帮扶的查检监督力度,并进行通报。

二、加强基础设施建设扶贫

建设和完善基础设施,加速交通、水利工程等基础设施的建设。对那些处于确实不适合人类居住恶劣生态环境中的特困少数民族群众,在自愿的原则下，采取搬迁式扶贫,但在实施这个浩大的工程中，要让这些搬迁移民搬得出、留得住、能就业有保障。

三、实施精准扶贫

用现代科学管理方法，统计和完善贫困人口信息管理系统，做到精准扶贫。要注重科学发展,从实际出发,实施精准扶贫。针对贫困的不同程度分门别类,分为一类相对贫困村和一类相对贫困人口,二类相对贫困村和二类相对贫困人口,对前者适当予以重点帮扶,在资金、项目发展方面给予倾斜,区别对待,把钱用在刀刃上。

四、加大“造血”式扶贫

要做好教育和引导工作。扶贫不仅要在经济上帮扶,同时也要在精神、智力上给予帮扶。改变贫困村和贫困户的等、靠、要思想,变“输血”为“造血”,实行产业扶贫，认真做好科学的调查、实验、评估、分析工作,绝不能以牺牲环境为代价。

10 模型评价与推广

10.1 模型的优点

1. 本文中采用的模型是针对数据的特点和问题的要求,尽可能选择较为简单且适当的模型与统计方法进行针对分析，并且注重对模型的比较与检验。根据每个模型所对应的算法与图形特点，广泛利用 MATLAB、SPSS、Origin 和 Python 数据处理、编程与统计软件，以使结果更加科学化与可视化。

2. 对于本论文中的数据，其特殊性在于已进行标准化处理，需要在此基础上继续对数据的异常值进行处理。鉴于数据已转化为相对位置的度量，对其计算增值能够更好的反映扶贫绩效变化程度。对于评级的划分，利用 Logistic 计算概率的结果作为级别的划

分依据，具有一定的创新实用性。

10.2 模型的缺点

模型具有一定局限性，由于实际情况外部影响因素和不确定性因素较多，模型存在一定客观数量偏差。样本案例的数量巨大以及数据难确定性，给建立模型带来了一定的阻碍。

10.3 模型的改进与推广

数据的评价模型除了本文所应用的 Topsis 以及传统的统计和非参方法之外，还有模糊综合评价法，AHP 层次分析，灰色关联，灰色预测模型等，可以进行运用后互相检验。同时，对于本文所采用的扶贫绩效的模型，能够广泛应用于其他的绩效评价场合。

11 参考文献

[1] 林毅夫. 关于我国扶贫政策的几点建议[J]. 发展, 2005.

[2] 王玉娜. 基于因子分析和数据包络法的河北省精准扶贫绩效评估研究[D]. 2019.

[3] 杜永红. 大数据背景下精准扶贫绩效评估研究简[J]. 求实, 2018.

[4] 李洁. 河南省扶贫开发绩效评价和对策研究[D]. 郑州大学.

[5] 庄天慧, 张海霞, 余崇媛. 西南少数民族贫困县反贫困综合绩效模糊评价——以 10 个国家扶贫重点县为例[J]. 西北人口, 2012(03):89-93.

[6] 贾俊平. 统计学. 2 版. 北京：清华大学出版社，2006.

[7] [美]肯·布莱克等. 以 Excel 为决策工具的商务与经济统计. 北京：机械工业出版社.

[8] 薛薇 . SPSS 统计分析方及应用（第二版），电子工业出版社。

[9] 王松桂，陈敏，陈立革编著.线性统计模型[M].高等教育出版社，1999.(9).50~70.

[10] 曾五一主编.统计学概论[M].首都经贸大学出版社，2008.(5) .70~110.

[11] 王强军. 基于类比方法的软件早期成本估算研究与应用. 国防科技大学硕士学位论文, 2008.

[12] Kun Yang,Zhenyu Yu,Yi Luo. Analysis on driving factors of lake surface water temperature for major lakes in Yunnan-Guizhou Plateau[J]. Water Research,2020,184.

[13] 李晨, 张杨, 陈长生.Logistic 回归应用的常见问题及其注意事项[J].中国儿童保健杂志,2020,28(03):358-360.

12 附 录

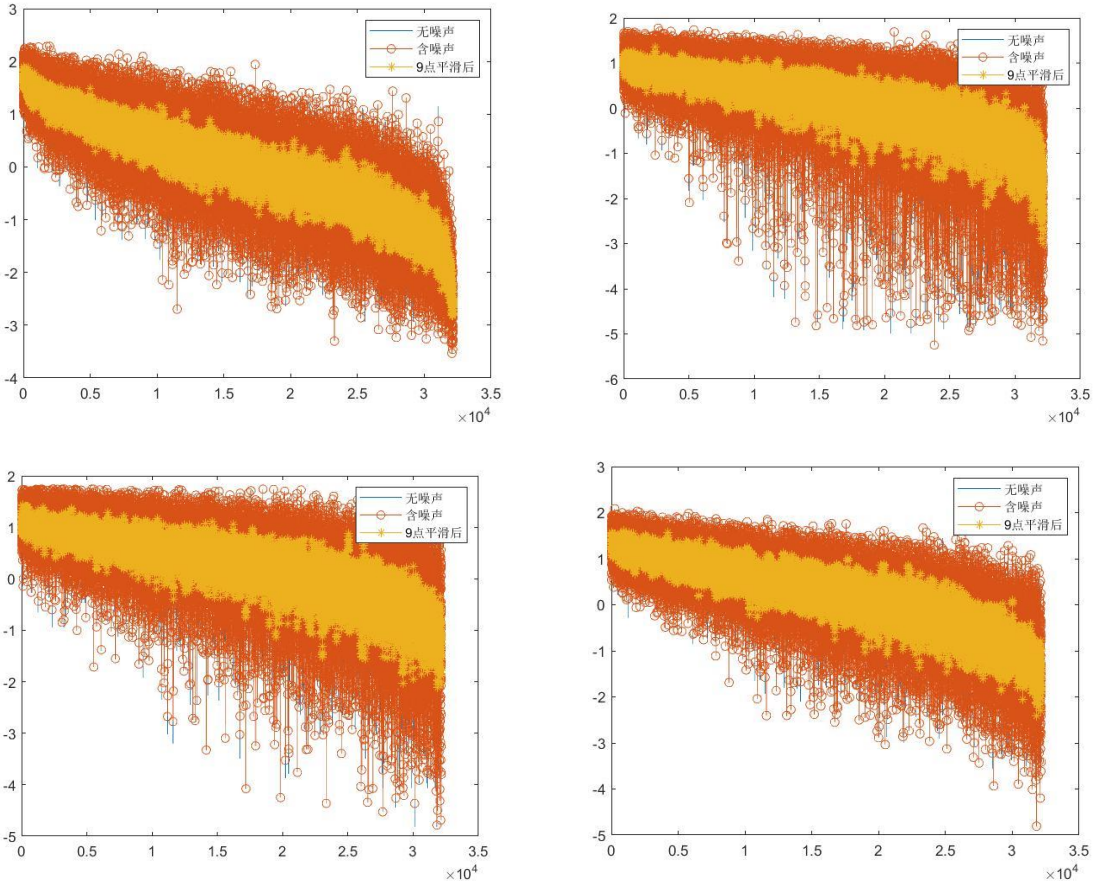
1. Matlab 数据平滑处理代码

```
function Y=smooth_data(y,n)
m=length(y);
j=1;
for i=(n-1)/2+1:(m-(n-1)/2)
    p=i-(n-1)/2;
    q=i+(n-1)/2;
   Y(j)=sum(y(p:q))/n;
   j=j+1;
end
end
t=x3';
n=length(t);
Y=t;
y=Y+(0.5-rand(1,n));
y3=smooth_data(y,9);
plot(1:n,Y,1:n,y,'-o',5:n-4,y3,'-*');
legend('无噪声','含噪声','9点平滑后');
```

2. 去除异常值代码

```
clc
clear
wt=xlsread('数据.xlsx);
[n,p]=size(wt);
 for c=1:p
  c_wt=wt(:,c);
  c_wtmean=mean(c_wt);
  j=1;
  for i=1:n
     vi(i,:)=c_wt(i,:)-c_wtmean;
     stdcwt=std(c_wt);
     if abs(vi(i,:))>3*stdcwt
          c_wt(i,:)=0;
          c_tbj(j,c)=i;
          j=j+1;
     end
  end
 ty_wt(:,c) = c_wt;
end
xlswrite('去异常后的数据',ty_wt);
```

3. 平滑处理前后数据对比图



4. Python3.7 Topsis 模型代码

```
import pandas as pd
data=pd.DataFrame({'x1':[-0.34,0.04,0.01,-0.13,-0.24,-0.25,-0.04,0.00,-0.24,-0.07,-0.15,-0.40,
0.11,-0.06,0.06,-0.02,0.06,-0.06,-0.17,-0.21,-0.45,-0.23,-0.14,-0.23,-0.07,-0.08,-0.19,-0.09,0.0
5,-0.40,-0.22,0.02,-0.10,-0.45,0.17,-0.27,-0.37,-0.22,-0.04,0.06,0.08,-0.16,-0.09,-0.15,-0.11,0.
07,-0.40,-0.52,-0.19,-0.19,0.01,-0.12,0.08,-0.33,0.49,-0.50,-0.37,-0.24,-0.36,-0.10,0.08,0.07,-0
.40,-0.11,-0.13,0.03,-0.06,0.11,0.03,-0.12,0.10,0.02,0.03,-0.72,0.28,-0.04,0.01,-0.06,-0.15,0.0
1,0.03,-0.20,-0.11,0.07,-0.28,-0.21,-0.22,-0.20,-0.36,0.01,0.07,-0.12,0.07,-0.10,-0.12,0.04,-0.0
8,0.02,0.07,-0.23,-0.15,-0.24,0.02,0.09,0.02,-0.16,-0.01,-0.24,0.16,-0.01,0.08,0.19,0.30,0.10,0
.30,0.63,0.14,-0.17,0.03,-0.10,0.05,-0.04,0.15,-0.01,0.17,-0.02,-0.14,-0.40,-0.20,-0.03,0.26,0.4
1,-0.06,0.26,0.42,0.01,-0.01,0.28,0.07,0.18,-0.19,0.07,0.22,0.18,0.15,-0.05,0.16,0.21,-0.01,0.0
2,0.21,0.17,-0.01,0.20,0.18,0.19,0.16,0.11,0.23,-0.13],

'x2':[-0.01,-0.01,-0.10,-0.27,-0.45,-0.39,-0.21,-0.21,-0.28,-0.24,-0.22,-0.25,-0.09,-0.16,-0.48,-
0.52,-0.23,-0.70,-0.93,-0.78,-0.95,-0.02,-0.03,-0.32,-0.42,-0.13,-0.39,-0.54,-0.23,-0.45,-0.92,-
0.37,-0.19,-0.15,-0.18,-0.48,-0.39,-0.46,-0.37,-0.42,-0.50,-0.13,-0.44,-0.64,-0.57,-0.58,-0.99,-
0.42,-0.48,-0.57,-0.36,-0.43,-0.49,-0.54,-0.46,-0.78,-0.52,-0.60,-0.45,-0.43,-0.48,-0.10,-0.72,-
0.21,-0.12,-0.30,-0.28,-0.16,-0.31,-0.13,-0.10,-0.09,0.08,-0.34,-0.24,-1.06,-0.76,-0.43,-0.08,0.
```

06,-0.06,-0.25,-0.09,-0.57,-0.11,-0.08,-0.57,-0.69,-0.67,0.02,-0.18,-0.25,-0.39,-0.55,-0.33,-0.39,-0.36,-0.43,-0.54,-0.05,-1.20,-0.42,-0.16,-0.46,-0.25,-0.17,-0.35,0.00,-0.24,-0.16,-0.28,-0.03,0.14,0.60,0.21,0.11,0.12,-0.08,-0.05,-0.34,0.05,-0.44,0.06,-0.08,-0.06,-0.45,-0.24,-0.31,-0.15,0.02,-0.03,-0.38,-0.64,0.08,-0.12,-0.04,-0.39,-0.07,-0.03,-0.12,-0.31,-0.31,-0.38,-0.61,-0.52,-0.69,0.09,-0.13,-0.02,-0.03,-0.05,-0.13,-0.13,-0.72,-0.03,-0.13,0.17,-0.14,-0.32,-0.19],

'x3':[0.32,0.02,-0.09,-0.22,-0.26,-0.15,-0.23,0.17,-0.37,-0.12,-0.15,-0.23,-0.03,-0.11,-0.56,-0.47,-0.50,-0.50,-0.97,-0.26,-0.59,0.21,0.07,-0.09,-0.26,0.08,-0.27,-0.28,0.03,-0.22,-0.11,-0.48,-0.03,0.15,-0.43,-0.47,-0.34,-0.22,-0.54,-0.10,-0.30,0.24,-0.24,-0.48,-0.36,-0.50,-1.01,-0.19,-0.17,-0.25,-0.51,-0.26,-0.26,-0.02,-0.50,-0.37,-0.38,-0.42,-0.42,-0.46,-1.10,-0.15,-0.78,-0.50,-0.21,-0.61,-1.26,-0.46,-0.35,-0.20,0.05,-0.47,-0.78,-0.48,-0.33,-1.68,-1.04,-0.64,-0.32,-0.30,-0.37,-0.21,-0.20,-0.56,-0.03,-0.20,-0.61,-0.49,-0.43,0.26,-0.15,-0.52,-0.20,-0.42,-0.09,-0.48,-0.05,-0.28,-0.64,0.03,-0.69,-0.62,-0.09,-0.46,-0.11,-0.15,-0.35,0.09,-0.17,0.19,-0.26,-0.02,-0.17,-0.59,-0.17,-0.32,0.19,-0.20,-0.31,-0.40,-0.62,-0.40,-0.14,-0.25,-0.08,-0.49,-0.25,-0.63,0.19,0.17,-0.03,-0.33,-0.79,-0.27,-0.39,-0.14,-0.31,-0.15,0.21,-0.04,-0.46,-0.26,-0.21,-0.37,-0.39,-0.40,0.15,-0.27,-0.01,-0.02,-0.21,-0.18,-0.37,-0.60,-0.08,0.13,-0.12,-0.12,-0.28,0.00],

'x4':[0.06,-0.08,-0.19,-0.57,-0.48,-0.57,-0.26,-0.36,-0.30,-0.40,-0.34,-0.54,-0.16,-0.36,-0.25,-0.46,-0.61,-0.31,-0.62,-0.49,-0.76,-0.11,-0.07,-0.33,-0.65,-0.17,-0.68,-0.53,-0.23,-0.42,-0.47,-0.54,-0.40,-0.14,-0.16,-0.68,-0.30,-0.55,-0.53,-0.26,-0.50,0.13,-0.38,-0.48,-0.49,-0.04,-0.35,-0.52,-0.61,-0.20,-0.51,-0.29,-0.04,-0.52,-0.22,-0.36,-0.78,-0.66,-0.57,-0.48,-0.84,0.05,-0.61,-0.40,-0.09,-0.23,-0.75,-0.26,-0.21,-0.54,-0.08,-0.02,-0.50,-0.44,-0.08,-0.78,-0.49,-0.24,-0.23,-0.28,-0.25,-0.41,-0.16,-0.38,-0.10,-0.57,-0.76,-0.62,-0.52,-0.17,-0.31,-0.15,-0.06,-0.31,-0.29,-0.30,-0.36,-0.39,-0.46,0.02,-0.87,-0.58,-0.16,-0.50,-0.35,-0.39,-0.66,-0.07,-0.43,0.04,-0.37,-0.25,-0.13,-0.44,0.20,0.26,0.20,-0.12,-0.32,-0.27,-0.36,-0.23,-0.17,-0.21,0.10,-0.45,-0.24,-0.22,0.04,0.07,0.20,0.03,-0.89,-0.05,0.15,-0.03,-0.32,-0.22,-0.12,-0.30,-0.62,-0.57,-0.58,-0.64,-0.55,-0.74,-0.06,-0.35,-0.37,0.14,-0.32,-0.12,-0.18,-0.54,-0.17,0.10,0.26,-0.35,0.01,-0.33],

'x5':[0.25,-0.06,-0.16,-0.39,-0.62,-0.59,-0.27,-0.05,-0.35,-0.24,-0.43,-0.28,-0.33,-0.45,-0.26,-0.75,-0.52,-0.53,-0.54,-0.39,-0.83,0.07,0.06,-0.42,-0.67,-0.13,-0.59,-0.75,-0.36,-0.74,-0.47,-0.67,-0.15,0.05,-0.45,-0.68,-0.49,-0.62,-0.74,-0.72,-0.88,0.04,-0.63,-0.48,-0.41,-0.07,-0.66,-0.52,-0.52,-0.66,-0.19,-0.23,-0.29,-0.50,-0.04,-0.18,-0.70,-0.63,-0.62,-0.79,-0.46,-0.05,-0.53,-0.50,-0.24,-0.29,-0.77,-0.55,-0.34,-0.44,-0.23,-0.35,-0.93,-0.56,-0.06,-0.65,-0.66,-0.68,-0.24,-0.06,-0.37,-0.13,0.00,-0.44,-0.06,-0.34,-0.62,-0.74,-0.63,0.07,-0.26,-0.24,-0.26,-0.11,-0.12,-0.09,-0.30,-0.58,-0.55,0.06,-0.68,-0.54,-0.09,-0.32,-0.43,-0.15,-0.15,0.16,-0.48,-0.02,-0.46,-0.29,-0.15,-0.51,0.16,0.09,0.10,-0.22,-0.24,-0.19,-0.04,-0.34,-0.05,-0.10,-0.04,-0.23,-0.30,-0.26,0.20,0.07,0.06,0.10,-0.81,0.17,-0.23,-0.13,-0.35,-0.19,-0.01,-0.27,-0.62,-0.41,-0.45,-0.62,-0.40,-0.88,0.07

,-0.26,-0.29,0.48,0.02,0.08,-0.21,-0.87,0.14,0.24,0.30,-0.13,-0.10,-0.11]})

```
import numpy as np
def entropyWeight(data):
    data = np.array(data)
    # 归一化
    P = data / data.sum(axis=0)
    # 计算熵值
    E = np.nansum(-P * np.log(P) / np.log(len(data)), axis=0)
    # 计算权系数
    return (1 - E) / (1 - E).sum()
entropyWeight(data)
def topsis(data,
weight=[0.39902052,0.0430893,0.09642889,0.15299766,0.01289316,0.13299988,0.0038237
1,0.00577422,0.00455109,0.00219075,0.1462308]):
    # 归一化
    data = data / np.sqrt((data ** 2).sum())
# 最优最劣方案
Z = pd.DataFrame([data.min(), data.max()], index=['负理想解', '正理想解'])
Result = data.copy()
Result['正理想解'] = np.sqrt(((data - Z.loc['正理想解']) ** 2 * weight).sum(axis=1))
Result['负理想解'] = np.sqrt(((data - Z.loc['负理想解']) ** 2 * weight).sum(axis=1))
Result['综合得分指数'] = Result['负理想解'] / (Result['负理想解'] + Result['正理想解'])
Result['排序'] = Result.rank(ascending=False)['综合得分指数']
Weight
    array([ 1.2897511 , -0.09997849, -0.06776056, -0.05919084, -0.06282122])
z
pd.set_option('display.max_rows',None)
Result
```

5. 二元 Logistic Python 代码

```
import numpy as np
import statsmodels.api as sm
import numpy as np
data=pd.DataFrame({'x1':[0.22,0.08,0.28,0.28,0.15,0.3,0.23,0.01,0.21,0.3,0.26,0.17,0.08,0.26,
0.16,0.07,0.21,0.18,0.07,0.19,0.09,0.11,0.17,0.18,0.19,0.17,0.07,0.06,0.06,0.08,0.08,0.07,0.06
,0.15,0.03,-0.06,0.02,0.11,-0.06,0.11,0.07,0.16,0.1,-0.05,0.02,0.04,0.03,-0.02,0.16,0.07,0.1,0.0
1,0.03,0.05,-0.15,0.14,0.05,-0.04,-0.02,0.03,0.07,-0.06,0.02,-0.01,-0.01,0.02,0.03,-0.06,-0.04,
```

0.07,0.02,-0.17,-0.07,-0.01,-0.09,0.04,0.01,0.01,-0.1,0.01,-0.01,-0.04,0,-0.09,-0.01,-0.11,-0.1,-0.06,-0.15,-0.11,-0.07,-0.1,-0.08,0.02,-0.01,0.01,-0.2,-0.12,-0.12,-0.13,-0.22,-0.12,-0.1,-0.19,-0.12,-0.21,-0.14,-0.19,-0.19,-0.08,-0.03,-0.19,-0.24,-0.15,-0.22,-0.24,-0.11,-0.13,-0.15,-0.22,-0.13,-0.27,-0.16,-0.24,-0.4,-0.2,-0.17,-0.24,-0.25,-0.21,-0.4,-0.23,-0.36,-0.45],

'x2':[-0.38,-0.48,-0.24,-0.07,-0.52,0.14,-0.32,-0.76,-0.13,0.21,0.08,-0.18,-0.5,-0.03,-0.24,-0.54,-0.05,-0.12,-0.57,-0.03,-0.46,-0.16,-0.13,-0.03,-0.13,-0.06,-0.58,-0.48,-0.23,-0.49,-0.28,-0.31,-0.42,0.06,0.08,-0.64,-0.37,-0.14,-0.28,-0.09,-0.39,0.09,0.6,-0.69,-0.43,-0.39,-0.3,-0.52,0.17,-0.18,-0.1,-0.36,-0.31,0.05,-1.2,0.12,-0.23,-0.37,-0.45,-0.06,-0.1,-0.7,-0.25,-0.35,-0.39,-0.09,-0.05,-0.43,-0.44,-0.03,-0.16,-0.93,-0.42,-0.13,-0.54,-0.01,-0.1,0.06,-0.43,-0.04,-0.08,-0.21,-0.21,-0.44,-0.02,-0.57,-0.55,-0.16,-0.64,-0.21,-0.24,-0.34,-0.36,-0.03,-0.16,0.02,-0.69,-0.43,-0.25,-0.27,-0.57,-0.13,-0.19,-0.31,-0.33,-0.78,-0.24,-0.57,-0.39,-0.13,0.02,-0.48,-0.6,-0.22,-0.92,-0.42,-0.09,-0.12,-0.08,-0.46,-0.19,-0.48,-0.17,-0.45,-0.99,-0.25,-0.08,-0.28,-0.39,-0.08,-0.72,-0.32,-0.67,-0.95],

'x3':[-0.21,-1.1,-0.33,-0.15,-0.39,-0.17,-0.28,-1.04,-0.27,-0.17,-0.27,-0.43,-0.3,-0.03,-0.17,-0.64,-0.21,-0.04,-0.56,-0.02,-0.46,-0.46,-0.18,-0.08,0.13,-0.08,-0.5,-0.56,-0.5,-0.26,-0.26,-0.26,-0.1,-0.14,-0.78,-0.79,-0.48,-0.12,-1.26,-0.03,-0.2,0.15,-0.59,-0.4,-0.28,-0.48,-0.61,-0.47,-0.12,-0.15,0.05,-0.51,-0.35,-0.62,-0.69,0.19,0.03,-0.54,-0.49,-0.37,-0.15,-0.5,-0.11,-0.35,-0.31,-0.47,-0.31,-0.64,-0.4,0.21,-0.09,-0.97,-0.26,-0.37,-0.28,0.02,-0.09,-0.3,-0.46,-0.14,-0.25,-0.23,0.17,-0.24,-0.01,-0.36,-0.42,-0.11,-0.48,-0.5,-0.12,-0.4,-0.05,-0.02,0.19,0.26,-0.49,-0.26,-0.52,-0.22,-0.61,-0.2,-0.03,-0.46,-0.09,-0.26,-0.25,-0.25,-0.27,0.08,0.17,-0.17,-0.42,-0.15,-0.11,-0.62,-0.2,-0.21,-0.32,-0.22,0,0.4,-0.15,-0.26,-1.01,-0.21,-0.2,-0.37,-0.15,-0.2,-0.78,-0.09,-0.43,-0.59],

'x4':[-0.58,-0.84,-0.08,-0.22,-0.55,-0.13,0.01,-0.49,-0.35,0.2,-0.05,-0.16,-0.5,0.2,-0.43,-0.46,-0.32,-0.3,-0.38,-0.25,-0.5,-0.26,-0.12,-0.17,0.1,0.1,-0.04,-0.25,-0.61,-0.04,-0.37,-0.57,-0.26,-0.17,-0.5,-0.89,-0.54,-0.35,-0.75,-0.16,-0.06,-0.06,-0.44,-0.74,-0.39,-0.3,-0.23,-0.46,0.26,-0.31,-0.08,-0.51,-0.21,-0.36,-0.87,0.2,-0.23,-0.53,-0.45,-0.25,0.05,-0.31,-0.35,-0.66,-0.32,-0.02,-0.32,-0.24,-0.23,-0.12,-0.16,-0.62,-0.65,-0.18,-0.53,-0.08,-0.19,-0.28,-0.48,-0.03,-0.21,-0.26,-0.36,-0.38,-0.37,-0.49,-0.31,-0.36,-0.48,-0.4,-0.4,-0.27,-0.36,0.14,0.04,-0.17,-0.62,-0.29,-0.15,-0.57,-0.76,-0.54,-0.4,-0.62,-0.29,-0.49,-0.24,-0.2,-0.68,-0.17,0.07,-0.61,-0.66,-0.34,-0.47,-0.58,-0.16,-0.09,-0.23,-0.55,-0.33,-0.68,-0.39,-0.48,-0.35,-0.41,-0.12,-0.3,-0.57,-0.57,-0.61,-0.33,-0.52,-0.76],

'x5':[-0.45,-0.46,-0.06,-0.19,-0.4,-0.15,-0.1,-0.66,-0.26,0.16,0.17,-0.45,-0.88,0.06,-0.48,-0.55,0.02,-0.27,-0.44,-0.29,-0.32,-0.55,0.08,0.14,0.24,-0.04,-0.07,-0.26,-0.52,-0.29,-0.46,-0.41,-0.72,-0.05,-0.93,-0.81,-0.67,-0.13,-0.77,-0.33,-0.26,0.07,-0.51,-0.88,-0.58,-0.09,-0.29,-0.75,0.3,-0.26,-0.23,-0.19,-0.34,-0.04,-0.68,0.1,-0.36,-0.74,-0.23,-0.37,-0.05,-0.53,-0.43,-0.15,-0.35,-0.35,-0.24,-0.68,-0.34,-0.01,-0.09,-0.54,-0.67,-0.21,-0.75,-0.06,-0.16,-0.06,-0.79,-0.13,-0.1,-0.27,-0.05,-0.63,-0.29,-0.41,-0.11,-0.45,-0.48,-0.5,-0.24,-0.19,-0.3,0.48,-0.02,0.07,-0.74,-0.23,-0.24,-0.39,-0.62,-0.44,-0.15,-0.62,-0.12,-0.39,-0.3,-0.66,-0.59,-0.13,0.07,-0.52,-0.63,-0.43,-0.47,-0.54,0,-0.24,-0.24,-0.62,-0.11,-0.68,-0.15,-0.62,-0.66,-0.13,-0.22,-0.35,-0.59,-0.34,-0.53,-0.42,-0.63,0.83],
'y':[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]})

```
y=data["y"]
X=data.ix[:,["x1","x2","x3","x4","x5"]]
from sklearn.linear_model import LogisticRegression
mode=LogisticRegression(C=500)
mode.fit(X,y)
print(mode.coef_)
[[66.06734903 -6.9577211    -5.79272331 -1.82590618 -4.75552632]]
print(mode.intercept_)
[-9.79299052]
b=mode.coef_
a=mode.intercept_
result=mode.fit(X,y)
from sklearn.model_selection import train_test_split,cross_val_score
precisions=cross_val_score(mode,X,y,cv=5,scoring="precision")
print(np.mean(precisions))
0.8947368421052632
recalls=cross_val_score(mode,X,y,cv=5,scoring="recall")
print(np.mean(recalls))
0.825
from sklearn import metrics
confusion=metrics.confusion_matrix(y,y_pred)
confusion
array([[92,    1],[ 4, 37]], dtype=int64)
import matplotlib.pyplot as plt
plt.rcParams["font.sans-serif"]=["SimHei"]
plt.matshow(confusion)
plt.title("混淆矩阵")
plt.colorbar
plt.ylabel("预测")
plt.xlabel("实际")
plt.show
y_prob=mode.predict_proba(X)
fpr,tpr,threshold=metrics.roc_curve(y,y_pred)
auc=metrics.auc(fpr,tpr)
auc
0.9458431681091004
plt.figure()
lw = 2
plt.plot(fpr, tpr, color='r')
plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
```

```
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.show()
print(metrics.classification_report(y,y_pred))
y_pred=mode.predict(X)
print(y_pred)
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```